PROBABILITY AND SAMPLING

Probability

Probability provides the language of data analysis.

Equiprobable outcomes definition:
Probability of event E is number of outcomes favorable to E divided by the total number of outcomes. e.g. Probability of a head = 1/2.

Long-run frequency definition:
If event E occurs n times in N identical experiments, the probability of E is the limit of n/N as N goes to infinity.

Subjective probability:
Probability is a measure of belief.

First Law of Probability

Law says that probability can take values only in the range zero to one and that an event which is certain has probability one.

$$\begin{cases} 0 \leq \text{Pr}(E) \leq 1 \\ \\ \text{Pr}(E|E) = 1 \text{ for any } E \end{cases}$$

i.e. If event E is true, then it has a probability of 1. For example:

$$\text{Pr(Seed is Round}|\text{Seed is Round)} = 1$$

Second Law of Probability

If G and H are mutually exclusive events, then:

$$\text{Pr}(G \text{ or } H) = \text{Pr}(G) + \text{Pr}(H)$$

For example,

$$\text{Pr}(\text{Seed is Round or Wrinkled}) = \text{Pr}(\text{Round}) + \text{Pr}(\text{Wrinkled})$$

More generally, if $E_i, i = 1, \ldots r$, are mutually exclusive then

$$\begin{aligned} \text{Pr}(E_1 \text{ or } \ldots \text{ or } E_r) &= \text{Pr}(E_1) + \ldots + \text{Pr}(E_r) \\ &= \sum_i \text{Pr}(E_i) \end{aligned}$$

Complementary Probability

If Pr(E) is the probability that E is true then Pr($\bar{E}$) denotes the probability that E is false. Because these two events are mutually exclusive

$$\text{Pr}(E \text{ or } \bar{E}) = \text{Pr}(E) + \text{Pr}(\bar{E})$$

and they are also exhaustive in that between them they cover all possibilities – one or other of them must be true. So,

$$\begin{aligned} \text{Pr}(E) + \text{Pr}(\bar{E}) &= 1 \\ \text{Pr}(\bar{E}) &= 1 - \text{Pr}(E) \end{aligned}$$

The probability that E is false is one minus the probability it is true.

Third Law of Probability

For any two events, G and H, the third law can be written:

$$\text{Pr}(G \text{ and } H) = \text{Pr}(G)\,\text{Pr}(H|G)$$

There is no reason why G should precede H and the law can also be written:

$$\text{Pr}(G \text{ and } H) = \text{Pr}(H)\,\text{Pr}(G|H)$$

For example

$$\begin{aligned}
& \text{Pr(Seed is round \& is type AA)} \\
= \; & \text{Pr(Seed is round|Seed is type AA)} \times \text{Pr(Seed is type AA)} \\
= \; & 1 \times p_A^2
\end{aligned}$$

Independent Events

If the information that H is true does nothing to change uncertainty about G, then

$$\text{Pr}(G|H) \;=\; \text{Pr}(G)$$

and

$$\text{Pr}(H \text{ and } G) \;=\; \text{Pr}(H)\,\text{Pr}(G)$$

Events G, H are independent.

Law of Total Probability

If $G, \bar{G}$ are two mutually exclusive and exhaustive events ($\bar{G} =$ not G), then for any other event E, the law of total probability states that

$$\Pr(E) = \Pr(E|G)\Pr(G) + \Pr(E|\bar{G})\Pr(\bar{G})$$

This generalizes to any set of mutually exclusive and exhaustive events $\{S_i\}$:

$$\Pr(E) = \sum_i \Pr(E|S_i)\Pr(S_i)$$

For example

$$\begin{aligned}
\Pr(\text{Seed is round}) &= \Pr(\text{Round}|\text{Type AA})\Pr(\text{Type AA}) \\
&\quad + \Pr(\text{Round}|\text{Type Aa})\Pr(\text{Type Aa}) \\
&\quad + \Pr(\text{Round}|\text{Type aa})\Pr(\text{Type aa}) \\
&= 1 \times p_A^2 + 1 \times 2p_A p_a + 0 \times p_a^2 \\
&= p_A(2 - p_A)
\end{aligned}$$

Bayes' Theorem

Bayes' theorem relates $\Pr(G|H)$ to $\Pr(H|G)$:

$$\begin{aligned}\Pr(G|H) &= \frac{\Pr(GH)}{\Pr(H)}, \text{ from third law} \\ &= \frac{\Pr(H|G)\Pr(G)}{\Pr(H)}, \text{ from third law}\end{aligned}$$

If $\{G_i\}$ are exhaustive and mutually exclusive, Bayes' theorem can be written as

$$\Pr(G_i|H) = \frac{\Pr(H|G_i)\Pr(G_i)}{\sum_i \Pr(H|G_i)\Pr(G_i)}$$

Bayes' Theorem Example

Suppose G is event that a man has genotype A_1A_2 and H is the event that he transmits allele A_1 to his child. Then $\Pr(H|G) = 0.5$.

Now what is the probability that a man has genotype A_1A_2 given that he transmits allele A_1 to his child?

$$\begin{aligned}
\Pr(G|H) &= \frac{\Pr(H|G)\Pr(G)}{\Pr(H)} \\
&= \frac{0.5 \times 2p_1p_2}{p_1} \\
&= p_2
\end{aligned}$$

Sampling

Statistical sampling: The variation among repeated samples from the same population ("fixed" sampling). Inferences can be made about that particular population.

Genetic sampling: The variation among replicate (conceptual) populations ("random" sampling). Inferences are made to all populations with the same history.

Classical Model

Reference population
(Usually assumed infinite and in equilibrium)

	↓		↓
Time 1	Population of size N	$\cdots$	Population of size N
	↓		↓
Time 2	Population of size N	$\cdots$	Population of size N
	↓		↓
	$\vdots$		$\vdots$
	↓		↓
Time t	Population of size N	$\cdots$	Population of size N

Sample of size n $\cdots$ Sample of size n

Coalescent Theory

An alternative framework works with genealogical history of a sample of alleles. There is a tree linking all alleles in a current sample to the “most recent common ancestral allele.” Allelic variation due to mutations since that ancestral allele.

The coalescent approach requires mutation and may be more appropriate for long-term evolution and analyses involving more than one species. The classical approach allows mutation but does not require it: within one species variation among populations may be due primarily to drift.